VegFru: A Domain-Specific Dataset for Fine-grained Visual Categorization

Saihui Hou, Yushan Feng and Zilei Wang
Department of Automation, University of Science and Technology of China
{saihui, fyushan}@mail.ustc.edu.cn, zlwang@ustc.edu.cn

Abstract

In this paper, we propose a novel domain-specific dataset named VegFru for fine-grained visual categorization (FGVC). While the existing datasets for FGVC are mainly focused on animal breeds or man-made objects with limited labelled data, VegFru is a larger dataset consisting of vegetables and fruits which are closely associated with the daily life of everyone. Aiming at domestic cooking and food management, VegFru categorizes vegetables and fruits according to their eating characteristics, and each image contains at least one edible part of vegetables or fruits with the same cooking usage. Particularly, all the images are labelled hierarchically. The current version covers vegetables and fruits of 25 upper-level categories and 292 subordinate classes. And it contains more than 160,000 images in total and at least 200 images for each subordinate class. Accompanying the dataset, we also propose an effective framework called HybridNet to exploit the label hierarchy for FGVC. Specifically, multiple granularity features are first extracted by dealing with the hierarchical labels separately. And then they are fused through explicit operation, e.g.*, Compact Bilinear Pooling, to form a unified representation for the ultimate recognition. The experimental results on the novel VegFru, the public FGVC-Aircraft and CUB-200-2011 indicate that HybridNet achieves one of the top performance on these datasets. The dataset and code are available at* https://github.com/ustc-vim/vegfru.

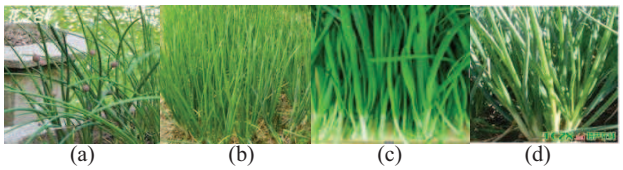
1. Introduction

In computer vision, fine-grained visual categorization (FGVC) refers to categorizing objects into subordinate classes, *e.g.*, breeds of birds or dogs. Compared to generic classification [6], FGVC needs to handle more subtle inter-class difference and larger intra-class variation of objects, thus requiring more discriminative and robust image representation. Recent years have witnessed the resurrection of deep convolutional neural network (DCNN), which holds state-of-the-art performance of various visual tasks [8, 25, 5]. The top-performing methods for FGVC [9, 36, 37] are also built upon DCNN and the training is data-hungry. However, the data with fine-grained labels is usually insufficient, *e.g.*, in CUB-200-2011 [28] there are only about 30 training images for each class. And the existing datasets for FGVC are mainly focused on domains of animal breeds, *e.g.*, birds [28] and dogs [11], or man-made objects, *e.g.*, cars [13] and aircrafts [18]. As the saying goes, *hunger breeds discontent*. In modern life, increasing attention has been paid to how to go on a balanced and nutritious diet. However, to the best of our knowledge, there is still no public dataset specially designed for recognizing the raw food materials and recommending appropriate recipes for individuals.

Figure 1. **Sample images in VegFru.** Top: vegetable images. Bottom: fruit images. Best viewed electronically.

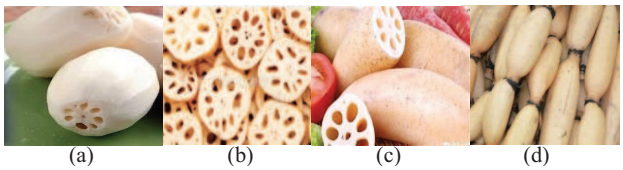
In this work, aiming at domestic cooking and food management, we introduce a novel domain-specific dataset named VegFru, which consists of vegetables and fruits that are closely associated with people's diet. In VegFru, vegetables and fruits are categorized according to their eating characteristics, *e.g.*, different edible parts of a certain vegetable or fruit, such as leaf and root, are classified into separate subordinate classes. And the objects in each image are the raw food materials, while the images that contain cooked food whose raw materials are indistinguishable are filtered out. Currently, the dataset covers vegetables and fruits of 25 upper-level categories and 292 subordinate classes[1], which has taken in all species in common. And it contains more than 160,000 images in total and at

[1]The upper-level category and subordinate class are respectively denoted as *sup-class* and *sub-class* in the following.



(a) (b) (c) (d)

Figure 2. **(a) chive (b) shallot (c) leek (d) green Chinese onion.** These images belong to different *sub-classes* but with subtle inter-class difference.



(a) (b) (c) (d)

Figure 3. **(a) lotus root (b) lotus root (c) lotus root (d) lotus root.** These images belong to the same *sub-classes* but with large intra-class variation.

least 200 images for each *sub-class*, which is much larger than the previous fine-grained datasets [11, 28, 13, 18]. Particularly, besides the fine-grained annotation, the images in VegFru are assigned hierarchical labels. And compared to the vegetable and fruit subsets of ImageNet [6], the taxonomy adopted by VegFru is more popular for domestic cooking and food management, and the image collection strictly serves this purpose, making each image in VegFru contain at least one edible part of vegetables or fruits with the same cooking usage. Some sample images are shown in Figure 1.

Our VegFru has the potential to be applied to the following aspects, but is not limited to them:

* *Fine-grained Visual Categorization (FGVC)*. The *sub-classes* in VegFru all belong to vegetables or fruits, and there exist subtle inter-class difference (Figure 2) and large intra-class variation (Figure 3). So it can be considered as a fine-grained dataset of novel domain, with more images available for each *sub-class*.

* *Hybrid-granularity methods for FGVC*. The label hierarchy has proved to be helpful for image recognition [33, 30, 32, 37]. With all images labelled hierarchically, VegFru is naturally well-suited for research on exploiting the hybrid-granularity information, *i.e.*, label hierarchy, for the challenging FGVC.

* *Practical applications for domestic cooking and food management*. VegFru collects enormous vegetable and fruit images of raw food materials and categorizes them by the eating characteristics. It is closely related to the daily life of everyone, and thus can promote the applications of computer vision in the Smart Home [4], *e.g.*, personalized recipe recommendation.

To verify the application value of VegFru, we also propose an effective framework named HybridNet with the aim of utilizing the hybrid-granularity information for FGVC, which aspect VegFru is well-suited for due to the label hierarchy. We take DCNN model to deal with the issue. In practice, DCNN is trained in a top-down manner, *i.e.*, the training is driven by the loss generated at the highest layer according to back propagation. When categorizing the coarse-grained *sup-classes*, the network only needs to handle generic attributes (*e.g.*, bird outline), but subtle characteristics (*e.g.*, bird eye, foot) become necessary when distinguishing the fine-grained *sub-classes*. Our HybridNet is exactly motivated by exploiting the complementarity between the coarse-grained and fine-grained features. Specifically, two-stream DCNNs are first trained by feeding the *sup-class* and *sub-class* labels separately, whose features are then fused to form a unified representation for the ultimate recognition. As for the fusion method, we adopt the advanced Compact Bilinear Pooling proposed by [7], in which the model is currently the best for FGVC without utilizing parts or external data. The experimental results on VegFru, FGVC-Aircraft [18] and CUB-200-2011 [28] show the robustness and superiority of HybridNet.

In summary, the main contributions of this work lie in two folds: VegFru and HybridNet. Specifically, the contribution of VegFru is highlighted in four aspects: novel domain, large scale, label hierarchy and application prospects. And HybridNet outperforms the model in [7] and achieves one of the top performance on the three datasets by exploiting the label hierarchy.

The rest of the paper is organized as follows. In Section 2, we introduce the construction of VegFru and perform detailed comparison of VegFru with the vegetable and fruit subsets of ImageNet and the existing fine-grained datasets. HybridNet and its related works are presented in Section 3. In Section 4, we set baselines on VegFru and experimentally evaluate the proposed HybridNet. Finally, the whole work is concluded in Section 5.

2. VegFru

2.1. Overview

We build the hierarchical structure of VegFru in accordance with the official literatures [21, 38]. Specifically, the vegetable hierarchy is constructed according to the Agricultural Biological Taxonomy described in [21][2], which is the most reasonable for the cooking purpose and arranges vegetables into root vegetable, cabbage, leafy vegetable, *etc.* Consequently, we obtain 15 *sup-classes* vegetables with 200 *sub-classes*. For fruits, similarly, we adopt the Horticultural Taxonomy in [38] to organize fruits into 10 *sup-*

[2]In fact, three are three taxonomies for vegetables in [21]. Besides the Agricultural Biological Taxonomy, the Botanical Taxonomy divides vegetables into two categories, namely monocotyledon and dicotyledon, and the Edible Organ Taxonomy groups vegetables into five categories, *i.e.*, root, stem, leaf, flower, and fruit.

Table 1. **The structure of VegFru.** **#Sub**-the number of *sub-classes* included. Perennial*-Perennial-miscellaneous vegetable. Persimmons*-Persimmons-jujubes fruit.

Sup-class	#Sub	Sup-class	#Sub
Aquatic vegetable	13	Alliaceous	10
Brassia oleracea	9	Beans	15
Bud seedling	4	Cabbage	5
Green-leafy vegetable	31	Eggplant	7
Perennial*	13	Melon	14
Tuber vegetable	10	Mushroom	24
Wild vegetable	32	Mustard	2
Root vegetable	11 (beetroot, black salsify, burdock root, carrot, celeriac, green radish, kohlrabi, parsnip, red radish, wasabi, white radish)		
Total	**15 *sup-classes* and 200 *sub-classes* for Veg200**		
Berry fruit	22	Drupe	13
Citrus fruit	13	Litchies	3
Persimmons*	6	Nut fruit	11
Pome	11	Other fruit	2
Collective fruit	5 (breadfruit, pineapple, sweetsop, annona muricata, artocarpus heterophyllus)		
Cucurbites	6 (golden melon, muskmelon, honey dew melon, papaya, netted melon, Hami melon)		
Total	**10 *sup-classes* and 92 *sub-classes* for Fru92**		

classes and 92 *sub-classes*. In the current version, there are 91,117 images for vegetables and 69,614 images for fruits. The number of images for each *sub-class* varies from 200 to 2000.

VegFru can be naturally divided into two subsets, *i.e.*, Veg200 for vegetables and Fru92 for fruits. Table 1 shows the structure of VegFru, where the *sup-classes* of Veg200 and Fru92 are listed along with the corresponding number of *sub-classes* included. And the *sub-classes* of Root vegetable, Collective fruit and Cucurbites are also listed[3].

2.2. VegFru Details

This section presents the details of VegFru. Specifically, we will respectively introduce the principles for building VegFru, process of collecting images, and dataset splits for training and test.

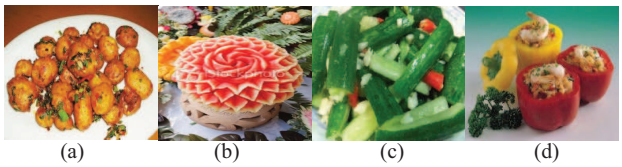
2.2.1 Building Principles

Aiming at domestic cooking and food management, VegFru is constructed according to the following principles.

[3]The detailed *sub-classes* for each *sup-class* are provided in the supplementary material.

(a) (b) (c) (d)

Figure 4. **(a) soybean (b) soybean (c) soybean seed (d) soybean seed.** Although (a)-(d) all belong to the seeds of soybean (in different growth periods), they are cooked in disparate ways, thus being classified into separate *sub-classes*.



(a) (b) (c) (d)

Figure 5. **(a) potato (b) watermelon (c) cucumber (d) pimento.** These images are dropped because the raw food materials are almost indistinguishable.

* The objects in the images of each *sub-class* have the same cooking usage. (Figure 3)
* Each image contains at least one edible part of vegetables or fruits. (Figure 1)
* The images that contain different edible parts of a certain vegetable or fruit, *e.g.*, leaf, flower, stem, root, are classified into separate *sub-classes*.
* Even for the images that contain the same edible part of given vegetable or fruit, if the objects are different in cooking, we also classify them into different *sub-classes*. (Figure 4)
* The objects in each image should be the raw food materials. If the raw materials of cooked food can not be made out, the images will be removed. (Figure 5)

2.2.2 Collecting Images

The above principles guide the construction of VegFru, as well as the process of image collection, which is a really challenging project.

The first step is to collect candidate images for each *sub-class*. The images are obtained by searching on the Internet, which is widely used to generate ImageNet [6] and Microsoft COCO [16]. The sources include Google, ImageNet, Flicker, Bing, Baidu, and so on. The retrieval keywords are the synonym sets of *sub-class* names in both Chinese and English. As a consequence, a large number of candidate images are collected. Specifically, over 800 images are gathered for each *sub-class*.

Then, to make the dataset highly reliable, the candidate images are further carefully processed through manual selection. In practice, the images of each *sub-class* are filtered

Figure 6. **Sample images of hyacinth bean.** Left: Two images without edible part in ImageNet; Right: Two images with edible part in VegFru.

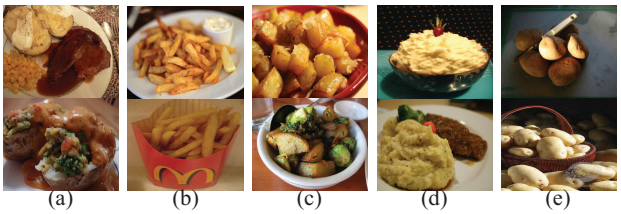


Figure 7. **Potatoes in the vegetable subsets of ImageNet.** Left To Right: (a) Baked potato (b) French fries (c) Home fries (d) Mashed potato (e) Uruguay potato. Only the images of Uruguay potato contain the raw food materials.

by ten people on the basis of the class description provided in [21, 38] and the true positives, following the principles described in Section 2.2.1. Only images affirmed by more than eight are reserved. The faded, binary, blurry and duplicated ones are all filtered out.

So far we have completed the construction of 25 *sup-classes* and 292 *sub-classes*, with more than 160, 000 images totally. Figure 1 displays some sample images collected by VegFru[4].

2.2.3 Dataset Splits

In VegFru, each *sub-class* contains at least 200 images, which are divided into training, validation and test set (denoted as *train*, *val* and *test* set in the following). An alternative split way is to first arrange the images in each *sub-class* randomly. Then the top 100 are selected for *train*, the following 50 for *val*, and the rest for *test*. Finally, a slight adjustment is applied to the split to ensure that each set is representative for the variability such as object numbers and background. The image list for each set is released attached in the dataset.

2.3. VegFru *vs*. ImageNet subsets

In this section, we compare VegFru with the vegetable and fruit subsets of ImageNet from three aspects, *i.e*., taxonomy, image selection and dataset structure. Through the comparison, the construction and usage of VegFru are further motivated.

Taxonomy. ImageNet [6] constructs its hierarchical structure based on WordNet [19], which organizes all the words according to the semantics. For vegetables and fruits, however, we tend to concentrate more on their eating characteristics in daily life. Actually the taxonomy adopted by ImageNet for vegetables and fruits is quite unpopular for domestic cooking and food management, and even contains many repeated categories. For example, turnip and radish simultaneously belong to root vegetable and cruciferous vegetable, and potato is grouped into root vegetable but is also on the list of solanaceous vegetable. In fact, according to [21], potato should be categorized into tuber vegetable. Moreover, some vegetables which are common in diet are not included in ImageNet, *e.g*., water spinach, shepherd's purse, basella rubra.

By contrast, in the construction of VegFru, we remove the rare categories, *e.g*. woad and ottelia, while many regular categories are added, *e.g*., Chinese pumpkin and sugarcane. And some categories are grouped into finer classes, *e.g*., radish are divided into white radish, red radish and green radish. The taxonomy adopted by VegFru specially serves the purpose of domestic cooking and food management in daily life.

Image Selection. All images in VegFru contain the edible part of a certain vegetable or fruit, which is not included in lots of images in ImageNet, as shown in Figure 6. Besides, some categories in ImageNet do not cover any raw food materials. For example, Figure 7 displays the images of five potato subordinate classes in the vegetable subsets of ImageNet, *i.e*., baked potato, French fries, home fries, mashed potato and Uruguay potato. Only Uruguay potato belongs to the raw food materials.

Dataset Structure. Table 2 shows some statistics of VegFru and the ImageNet subsets. In particular, there are 50 fruit *sub-classes* and 27 vegetable *sub-classes* whose number of images is less than 200 in ImageNet, while VegFru is comprised of 292 popular *sub-classes* of vegetables and fruits with more than 200 images for each *sub-classes*. And the taxonomy tree, *i.e*., the distribution of *sup-classes* and *sub-classes*, is reasonably reorganized for vegetables and fruits in VegFru.

Table 2. **VegFru *vs*. ImageNet subsets on dataset structure.** **#Sup**-the number of *sup-classes*. **#Sub**-the number of *sub-classes*. **Min/Max**-the minimum/maximum number of images in each *sub-class*. **#Sub<200**-the number of *sub-classes* that consist of less than 200 images.

	#Sup	#Sub	Min	Max	#Sub <200
Vegetables in ImageNet	25	175	3	1500+	27
Vegetables in VegFru	15	200	202	1807	0
Fruits in ImageNet	75	196	0	1500+	50
Fruits in VegFru	10	92	202	1615	0

[4]More sample images are provided in the supplementary material.

Table 3. **VegFru *vs*. Fine-grained Datasets. #Sup**-the number of *sup-classes*. **#Sub**-the number of *sub-classes*. **#Image**-the number of images in total. **#Train/#Val/#Test**-the number of images in *train/val/test* set. **#Train+Val(avg)**-the average number of images in each *sub-classes* for model training (include *train* and *val* set).

Dataset	#Sup	#Sub	#Image	#Train	#Val	#Train+ Val(avg)	#Test
Birds	none	200	11788	5994	none	~30	5794
Dogs	none	120	20580	12000	none	100	8580
Cars	none	196	16185	8144	none	~42	8041
Aircrafts	70	100	10000	3334	3333	~67	3333
VegFru	25	292	160731	29200	14600	150	116931

Figure 8. **Sample images in FGVC-Aircraft.** The airplanes occupy a large fraction of the whole images. And there exists only one airplane in each image with relatively clean background.

2.4. VegFru *vs*. Fine-grained Datasets

In this section we further compare VegFru with four representative fine-gained datasets, *i.e*., CUB-200-2011 [28] (Birds), Stanford Dogs [11] (Dogs), Stanford Cars [13] (Cars) and FGVC-Aircraft [18] (Aircrafts)[5], which are widely used in previous works [35, 12, 17, 7]. The detailed comparison is shown in Table 3. More fine-grained datasets, such as Oxford Flowers [20] and Pets [23], are not listed here out of the consideration of simplicity.

Compare to these existing datasets, the domain of VegFru is novel and more associated with people's daily life, which contributes to its broad application prospects. And VegFru is larger in scale, which has up to 150 images available in each *sub-classes* for model training. Particularly, all the images in VegFru are hierarchically categorized into *sup-classes* and *sub-classes*, while the images in these datasets, except FGVC-Aircraft, are only assigned with fine-grained labels. So VegFru is well-suited for the hybrid-granularity research on FGVC. We noticed that the previous works [30, 39] declared to annotate some fine-grained datasets, *e.g*., CUB-200-2011, with extra labels to get the label hierarchy. However, to the best of our knowledge, the annotation is not publicly available until the submission, and the labelling process is labor-intensive. Furthermore, though FGVC-Aircraft is with hierarchical labels, the objects of interest, *i.e*., aircrafts, usually occupy a large fraction of the whole images, and each image only contains one aircraft with relatively clean background (Figure 8). In contrast, the images in VegFru are with cluttered background and vary in number and scale of the objects (Figure 1).

[5]For FGVC-Aircraft, airplane variants are chosen as the labels of *sub-classes*, and the 70 *sup-classes* in Tabel 3 is the number of airplane families, which are the upper-level annotaions of airplane variants. Please refer to [18] for details of this dataset.

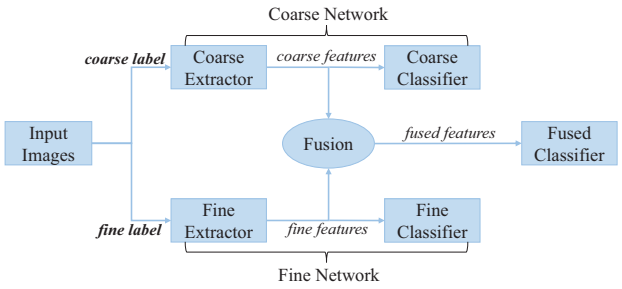


Figure 9. **Illustration of the proposed HybridNet.** Two-stream features which deal with the hierarchical labels are first extracted separately, and then sent through the *Fusion* module to train the *Fused Classifier* for overall classification.

3. HybridNet

Accompanying the dataset, we also propose an effective framework called HybridNet to conduct the image classification, illustrated in Figure 9. The motivation is to exploit the label hierarchy for FGVC, which can further verify the application value of VegFru.

Specifically, the input images with *sup-class* and *sub-class* labels (denoted as *coarse label* and *fine label* in Figure 9) are firstly sent into two DCNNs for separate classification. Here an end-to-end DCNN is logically divided into two functional parts, *i.e*., feature extractor and image classifier. The division can theoretically occur at any layer, *e.g*., *pool5* in VGGNet. Secondly, the features output by each extractor, *i.e*., *coarse features* and *fine features*, are sent through the *Fusion* module to form a unified representation, *i.e*., *fused features*. The advanced Compact Bilinear Pooling [7] is chosen as the fusion method. Finally, the *Fused Classifier* plays as the key component to aggregate two-level features for the ultimate recognition. Actually the *Fused Classifier* can handle either coarse-grained or fine-grained categorization, and the latter one which is more challenging is evaluated in our experiments. The training strategy of HybridNet will be elaborated in Section 4.2.1.

The design of HybridNet comes from the following philosophy. Since DCNN is trained in a top-down manner, the *coarse features* and *fine features* tend to deal with different aspects of the objects, with the condition of being fed with the *coarse label* and *fine label* separately. After the *Fusion*, the *fused features* has synthesized the hybrid-granularity information, so it is expected to be richer and more accurate than the *fine features*, thus resulting in higher accuracy for FGVC. From another perspective, the optimization of HybridNet is comprised of three tasks, *i.e*.,

* Categorizing *sup-classes* according to the *coarse features* in the *Coarse Classifier*

* Categorizing *sub-classes* according to the *fine features* in the *Fine Classifier*

* Categorizing *sub-classes* according to the *fused features* in the *Fused Classifier*.

Such multi-task optimization is beneficial to learn more discriminative image representation [22, 37, 34].

In the existing literatures, there are considerable interests to enhance DCNN with greater capacity for FGVC, *e.g.*, leveraging parts of objects [35, 30, 9, 36], embedding distance metric learning [24, 29, 31, 37]. Since HybridNet is motivated by exploiting the label hierarchy, here we only focus on the related works [30, 37, 32] that make use of label hierarchy for FGVC and clarify their differences with our work. In [30], the features of multiple granularity are concatenated before the linear SVM, whose training is independent from DCNN. While in HybridNet, the Compact Bilinear Pooling are adopted as fusion method and the model is trained in an end-to-end manner. In [37], the label hierarchy is used to construct the input triplets for jointly optimizing both classification and similarity constraints. In our opinions, the hybrid-granularity information is not fully utilized in this way. Our HybridNet shares similar ideas with [32]. However, in [32], the training set is augmented by the external data annotated with hyper-classes, while the images in original dataset are still only with fine-grained labels. In contrast, HybridNet is applied to the input images that are annotated with hierarchical labels, *e.g.*, VegFru, and the multiple granularity features are separately learned and fused through explicit operation.

Furthermore, the architecture of HybridNet is intuitively similar to the Bilinear CNN in [17], where the Bilinear Pooling is first proposed to aggregate the two-stream DCNN features for FGVC. Actually they differ from each other in three aspects. Firstly and most importantly, in [17], the two-stream DCNNs both deal with fine-grained categorization and much efforts are taken to break the symmetry of two networks. But in HybridNet, the two DCNNs are naturally asymmetric and complementary since they are fed with the coarse-grained and fine-grained labels separately. Secondly, the network architectures are actually dissimilar. The Bilinear CNN is eventually implemented by a single DCNN due to weight sharing, while HybridNet holds two DCNNs which do not share weights. Thirdly, the training process is quite different. Compared to single-task optimization of the Bilinear CNN, the training of HybridNet is made up of multiple tasks. In practice, we adopt the Compact Bilinear Pooling [7] as fusion method, which inherits the discriminative power of the Bilinear Pooling and meanwhile reduces the computation cost. The model in [7] is denoted as CBP-CNN in the following.

Table 4. **Baselines on VegFru.** The typical CaffeNet, VGGNet and GoogLeNet are chosen to set benchmarks on VegFru. All results are evaluated on the *test* set and reported in the top-1 mean accuracy.

Dataset	Category	CaffeNet	VGGNet	GoogLeNet
Veg200	15 *sup-classes*	74.92%	83.81%	83.50%
	200 *sub-classes*	67.21%	78.50%	80.17%
Fru92	10 *sup-classes*	79.86%	86.81%	87.54%
	92 *sub-classes*	71.60%	79.80%	81.79%
VegFru	25 *sup-classes*	72.87%	82.45%	82.52%
	292 *sub-classes*	66.40%	77.12%	79.22%

4. Experiment

In the experiments, we first set benchmarks on VegFru, and then compare HybridNet with the corresponding baselines on VegFru, FGVC-Aircraft [18] and CUB-200-2011 [28]. All the networks are implemented with Caffe [10].

4.1. VegFru Baselines

4.1.1 Experimental Setup

The choice of features is usually treated as the most important design in image recognition, and so far DCNN is considered to be the most competitive method for feature extraction. To comprehensively evaluate VegFru, therefore, we adopt the representative DCNN architectures including CaffeNet [14], VGGNet [26], and GoogLeNet [27] (available in the Caffe Model Zoo [1]) to set benchmarks.

All the networks are pretrained on ImageNet and then finetuned on VegFru. The images are randomly flipped before passing into the networks and no other data augmentation is used. The base learning rate is set to 0.001 and reduced by a factor of 10 when the loss plateaus. The test is done with one center crop of the input images. Finally the top-1 mean accuracy is taken to measure the classification performance. It is worth mentioning that the dataset split way follows the description in Section 2.2.3. The *train* set is used for the finetuning and the evaluation is performed on the *test* set. The *val* set is taken for error analysis here[6].

4.1.2 Quantitative Results

The experiments are carried on *sup-classes* and *sub-classes* of VegFru as well as its subsets, *i.e.*, Veg200 and Fru92, and the quantitative results are shown in Table 4. The three networks all achieve reasonable performance for the task of image classification, which validates the reliability of VegFru. However, even the best top-1 accuracy with GoogLeNet

[6]The top-1 mean accuracy on *val* set with CaffeNet, VGGNet, and GoogLeNet is provided in the supplementary material.

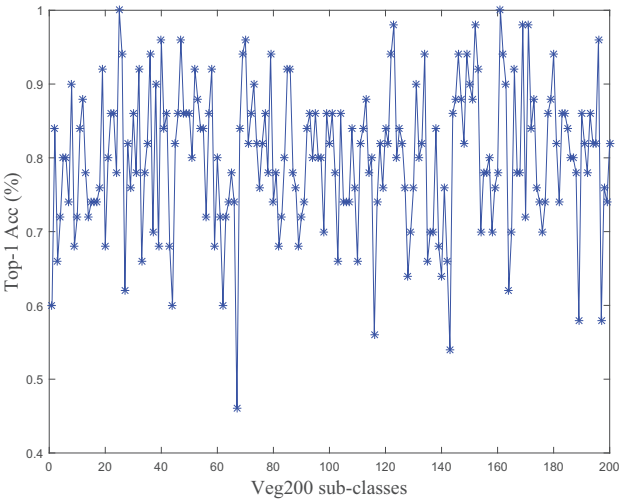


Figure 10. **The top-1 accuracy of GoogLeNet on each *sub-classes* of Veg200.** The results are evaluated on the *val* set and the lowest accuracy lies in the 67th *sub-class*, *i.e.*, dandelion.

Figure 11. **Sample images of dandelion**

Figure 12. **Left: shepherd's purse; Right: prickly lettuce.**

(79.22% on *sub-classes* of VegFru) is still not satisfying enough for real-world applications, indicating that it is still vital and necessary to develop more advanced models for the recognition.

4.1.3 Error Analysis

Along with reporting the top-1 mean accuracy, we also analyze the classification performance on each *sub-class*. Here GoogLeNet is taken to illustrate the proof and evaluated on Veg200. The *val* set is chosen for the evaluation since it has equal number of images for each *sub-classes*. The analysis results are shown in Figure 10, and the lowest accuracy (46%) lies in the *sub-class* of dandelion (Figure 11). We further look into the result and find that lots of misclassified images are predicted to be shepherd's purse and prickly lettuce (Figure 12). It can be seen that the images in Figure 11 and Figure 12 are of subtle difference, and thus more robust image representation is required to discriminate them.

4.2. HybridNet Performance

4.2.1 Implementation Details

The DCNN in HybridNet can be any existing model, *e.g.*, CaffeNet, VGGNet or GoogLeNet. Here the 16-layer VGGNet [26] is selected to construct the HybridNet as in CBP-CNN [7]. Specifically, the feature extractor of HybridNet is comprised of the layers of VGGNet before *pool5*. And the image classifier includes the layers of *compact bilinear pooling*, *signed square-root*, l_2*-normalization*, *fully-connection* and *softmax*.

The *Coarse Network* and *Fine Network* in HybridNet are the variants of CBP-CNN. So before introducing the training of HybridNet, we first review the training process of CBP-CNN, which consists of two stages denoted as ***ft_last_layer*** and ***ft_all*** [2]. Specifically, ***ft_last_layer*** is used to train the layers after the Compact Bilinear Pooling starting with a high learning rate (*e.g.*, 1), and ***ft_all*** means global finetuning with a relatively low learning rate (*e.g.*, 0.001). The training strategy of HybridNet is illustrated in Figure 13. Firstly, the *Coarse Network* and *Fine Network* are trained in parallel by feeding the *coarse label* and *fine label* separately (each including ***ft_last_layer*** and ***ft_all*** shown in Figure 13(a)(b)). Secondly, the *Fused Classifier* is optimized based on the *fused features* with the rest fixed (Figure 13(c)). Finally, the whole network is globally finetuned (Figure 13(d)). The *Coarse Classifier* and *Fine Classifier* are removed in the global finetuning (Figure 13(d)), since the jointly finetuning strategy does not help in this case, which will be further discussed in Section 4.2.3. The detailed parameters for the training and test is released with the dataset and code.

4.2.2 Performance Comparison

The baseline of HybridNet is set by replacing the *coarse label* with the *fine label* in Figure 9. In that condition, the two DCNNs with the same architectures are symmetrically initialized and remain symmetric after finetuning since the gradients for two networks are identical [17]. Thus the model can be implemented with just a singe DCNN. So it is natural to treat CBP-CNN as the baseline of HybridNet.

The performance comparison for HybridNet on VegFru, FGVC-Aircraft and CUB-200-2011 is shown in Table 5, where the results are reported in the top-1 mean accuracy. For HybridNet, the output of *Fused Classifier* is taken for the evaluation. As far as we know, CBP-CNN is the existing state-of-the-art method for FGVC without utilizing parts or external data. Our HybridNet outperforms CBP-CNN by more than 1.3% on VegFru and FGVC-Aircraft, which both contain hierarchical labels. For CUB-200-2011, extra efforts are first taken to construct the label hierarchy according to the taxonomy in North American Birds [3], and

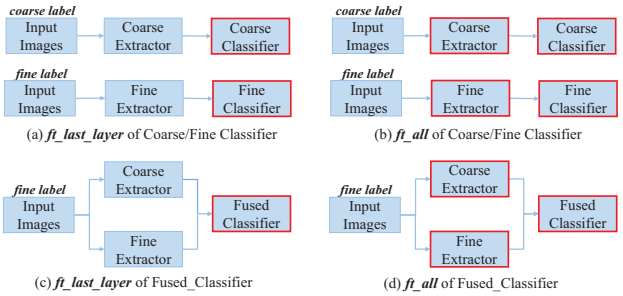


Figure 13. **Training strategy of HybridNet.** Inspired by the training of CBP-CNN, the training of the *Fused Classifier* is also divided into two stages following the same denotations ((c)(d)). The *Fusion* moduels in (c)(d) are omitted for simplicity. In each stage, only the components surrounded by the red rectangle are finetuned with the rest fixed. Best viewed electronically.

Table 5. **Performance comparison for HybridNet.** To keep the experiments consistent, HybridNet is trained on the *train* set of VegFru. And it is trained on the *trainval* set of FGVC-Aircraft [18] and *train* set of CUB-200-2011 [28]. Finally, all results are evaluated on the *test* set and reported in the top-1 mean accuracy.

Dataset	VegFru (292 *sub-classes*)	Aircrafts [18] (100 *sub-classes*)	CUB [28] (200 *sub-classes*)
VGGNet [26]	77.12%	84.46%	72.32%
CBP-CNN [7]	82.21%	87.49%	84.91%
HybridNet (ours)	**83.51%**	**88.84%**	**85.78%**

then HybridNet is applied to obtain 85.78% on this dataset which is higher than that with CBP-CNN (84.91%). The improvement on CUB-200-2011 is less significant, which is probably due to the small size of training set. The experimental results indicate that the label hierarchy does help for FGVC[7].

4.2.3 Discussion

In the global finetuning of HybridNet (Figure 13 (d)), we have tried to add the *Coarse Classifier* and *Fine Classifier* as regulations, for the sake of making the features separately learned by each extractor discriminative alone in the training process [15, 5]. However, our preliminary experiments indicate that this jointly finetuning strategy does not suit for this case and instead brings performance degradation, which is probably caused by the complexity of the Compact Bilinear Pooling for optimizing. Actually, it has been proved in [34] that the jointly finetuning strategy does not always work. The training strategy of HybridNet is equivalent to the iterative switchable learning scheme adopted in [34], *i.e.*, multiple tasks are optimized by turns (Figure 13 (a)-(d)). Besides, there still exists room to improve HybridNet, *e.g.*, the *Coarse Network* and *Fine Network* can share some shallow layers as the model in [32], which has the potential to reduce the GPU memory consumption.

[7]We also provide the results of evaluating HybridNet on the coarse-grained categorization in the supplementary material.

5. Conclusion

In this work, we construct a domain-specific dataset, namely VegFru, in the field of FGVC. The novelty of VegFru is that it aims at domestic cooking and food management, and categorizes vegetables and fruits according to their eating characteristics. In VegFru, there are at least 200 images for each subordinate class with hierarchical labels, and each image contains at least one edible part of vegetables or fruits with the same cooking usage. It is closely associated with the daily life of everyone and has broad application prospects. Besides, HybridNet is proposed accompanying the dataset to exploit the label hierarchy for FGVC. In HybridNet, multiple granularity features are first separately learned and then fused through explicit operation, *i.e.*, Compact Bilinear Pooling, to form a unified image representation for overall classification. The results on VegFru, FGVC-Aircraft and CUB-200-2011 demonstrate that HybridNet achieves one of the top performance on these datasets. We believe that our VegFru and HybridNet would inspire more advanced research on FGVC.

Acknowledgment

This work is supported partially by the National Natural Science Foundation of China under Grant 61673362 and 61233003, Youth Innovation Promotion Association CAS, and the Fundamental Research Funds for the Central Universities. Many thanks to Dequan Wang for offering the label hierarchy of CUB-200-2011. And we are grateful for the generous donation of Tesla GPU K40 from the NVIDIA corporation.

References

[1] https://github.com/BVLC/caffe/wiki/Model-Zoo. 6

[2] https://github.com/gy20073/compact_bilinear_pooling. 7

[3] https://birdsna.org/Species-Account/bna/home. 7

[4] M. Chan, D. Estève, C. Escriba, and E. Campo. A review of smart homes - present state and future challenges. *Computer Methods and Programs in Biomedicine*, 91:55–81, 2008. 2

[5] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *CVPR*, 2016. 1, 8

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1, 2, 3, 4

[7] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell. Compact bilinear pooling. In *CVPR*, 2016. 2, 5, 6, 7, 8

[8] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *ECCV*, 2016. 1

[9] S. Huang, Z. Xu, D. Tao, and Y. Zhang. Part-stacked cnn for fine-grained visual categorization. In *CVPR*, 2016. 1, 6

[10] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 6

[11] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *CVPR Workshop*, 2011. 1, 2, 5

[12] J. Krause, H. Jin, J. Yang, and L. Fei-Fei. Fine-grained recognition without part annotations. In *CVPR*, 2015. 5

[13] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV Workshop*, 2013. 1, 2, 5

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 6

[15] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. Deeply-supervised nets. In *AISTATS*, 2015. 8

[16] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 3

[17] T.-Y. Lin, A. RoyChowdhury, and S. Maji. Bilinear cnn models for fine-grained visual recognition. In *ICCV*, 2015. 5, 6, 7

[18] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013. 1, 2, 5, 6, 8

[19] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 4

[20] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, 2008. 5

[21] Chinese Academy of Agricultural Sciences. *Vegetable Cultivation (Second Edition, In Chinese)*. Beijing: China Agriculture Press, 2010. 2, 4

[22] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC*, 2015. 6

[23] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, 2012. 5

[24] Q. Qian, R. Jin, S. Zhu, and Y. Lin. Fine-grained visual categorization via multi-stage metric learning. In *CVPR*, 2015. 6

[25] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 1

[26] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6, 7, 8

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 6

[28] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 1, 2, 5, 6, 8

[29] C. Wah, G. V. Horn, S. Branson, S. Maji, P. Perona, and S. J. Belongie. Similarity comparisons for interactive fine-grained categorization. In *CVPR*, 2014. 6

[30] D. Wang, Z. Shen, J. Shao, W. Zhang, X. Xue, and Z. Zhang. Multiple granularity descriptors for fine-grained categorization. In *ICCV*, 2015. 2, 5, 6

[31] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. Learning fine-grained image similarity with deep ranking. In *CVPR*, 2014. 6

[32] S. Xie, T. Yang, X. Wang, and Y. Lin. Hyper-class augmented and regularized deep learning for fine-grained image classification. In *CVPR*, 2015. 2, 6, 8

[33] Z. Yan, H. Zhang, R. Piramuthu, V. Jagadeesh, D. DeCoste, W. Di, and Y. Yu. Hd-cnn: Hierarchical deep convolutional neural networks for large scale visual recognition. In *ICCV*, 2015. 2

[34] C. Zhang, H. Li, X. Wang, and X. Yang. Cross-scene crowd counting via deep convolutional neural networks. In *CVPR*, 2015. 6, 8

[35] N. Zhang, J. Donahue, R. B. Girshick, and T. Darrell. Part-based r-cnns for fine-grained category detection. In *ECCV*, 2014. 5, 6

[36] X. Zhang, H. Xiong, W. Zhou, W. Lin, and Q. Tian. Picking deep filter responses for fine-grained image recognition. In *CVPR*, 2016. 1, 6

[37] X. Zhang, F. Zhou, Y. Lin, and S. Zhang. Embedding label structures for fine-grained feature representation. In *CVPR*, 2016. 1, 2, 6

[38] Z. Zheng, S. Zhang, and Z. Zhang. *Fruit Cultivation (In Chinese)*. Beijing: Press of Agricultural Science and Technology of China, 2011. 2, 4

[39] F. Zhou and Y. Lin. Fine-grained image classification by exploring bipartite-graph labels. In *CVPR*, 2016. 5